



po daac

Physical Oceanography Distributed Active Archive Center



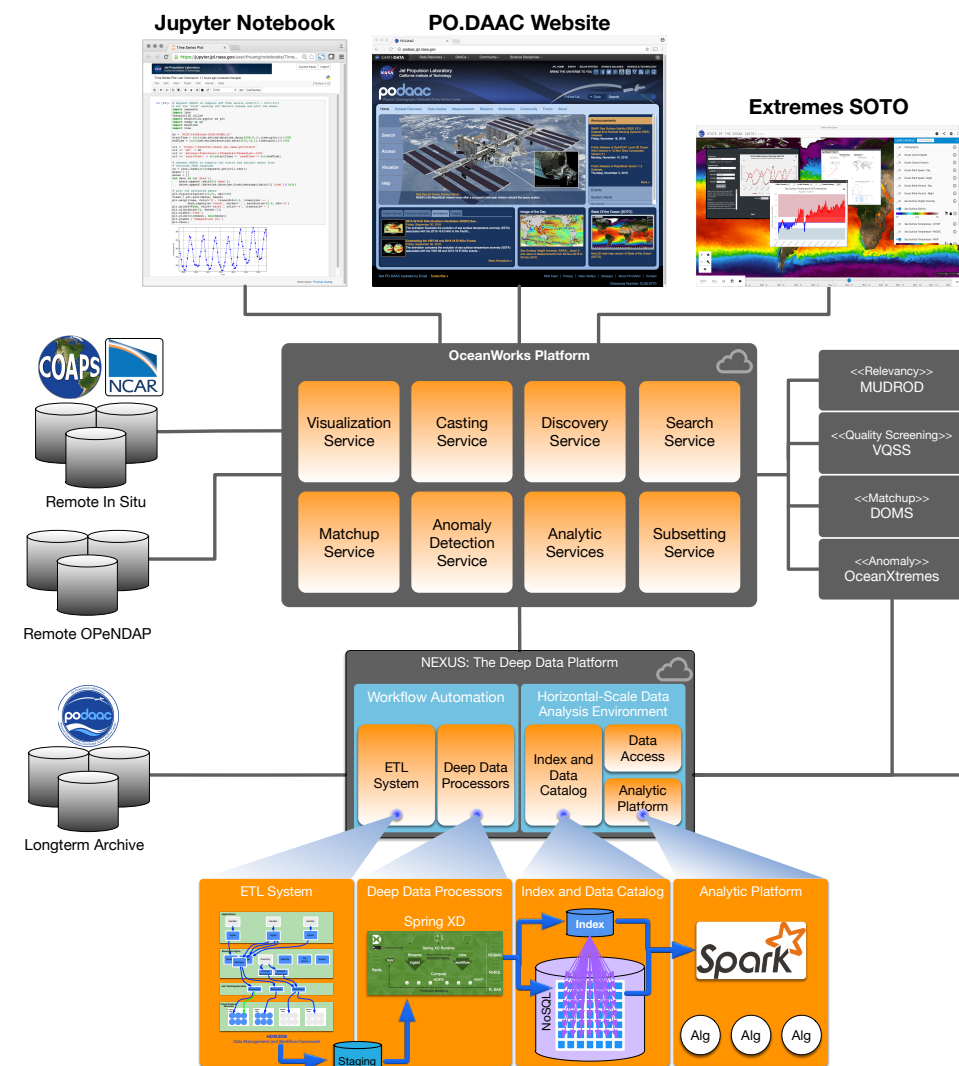
A tiled storage technology for ocean data access and analytics

Edward Armstrong¹, Thomas Huang¹ in conjunction with PO.DAAC and Oceanworks development team

1. Jet Propulsion Laboratory, California Institute of Technology

Oceanworks -- NASA AIST

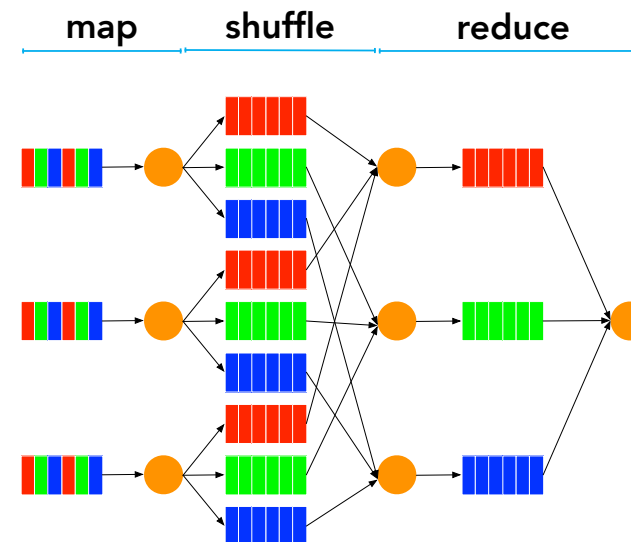
- Enabling Next Generation of Ocean Science Tools and Services with data analytics and services for
 - Oceanographic anomaly detection
 - Sea level rise and variation
 - Hydrological mass change
 - In situ to satellite matchups
 - Ocean models (in progress)
- The engine behind this capability is Nexus
 - A tiled storage system based on Casandra and Solr
 - Data analytics by Apache Spark
 - The entire software stack is an Apache Open Source project:
 - Integrated Ocean Science Data Analytics Platform (SDAP)
 - Capabilities accessible (e.g. in a Jupyter notebook) via RESTful APIs
- Key message: *harmonize data, tools, services and computation resources so the scientist can concentrate on results and not on data conditioning or preparation*



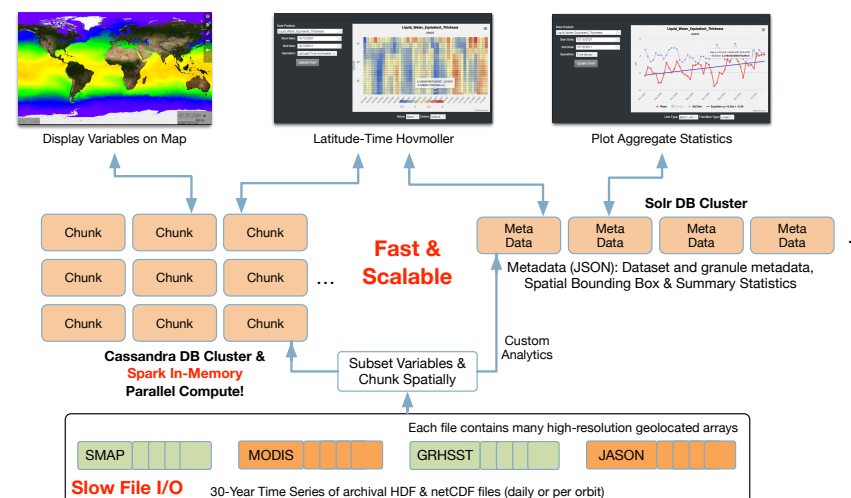
NASA AIST OceanWorks

Scalable Data Analytic Solution

- **MapReduce:** A programming model for expressing distributed computations on massive amount of data and an execution framework for large-scale data processing on clusters of commodity servers. - J. Lin and C. Dyer, “Data-Intensive Text Processing with MapReduce”
 - **Map:** splits processing across cluster of machines in parallel, each is responsible for a record of data
 - **Reduce:** combines the results from Map processes
- **SDAP’s NEXUS** is a data-intensive analysis solution using a new approach for handling science data to enable large-scale data analysis
 - Streaming architecture for horizontal scale data ingestion
 - Scales horizontally to handle massive amount of data in parallel
 - Provides high-performance geospatial and indexed search solution
 - Provides tiled data storage architecture to eliminate file I/O overhead
 - A growing collection of science analysis webservice



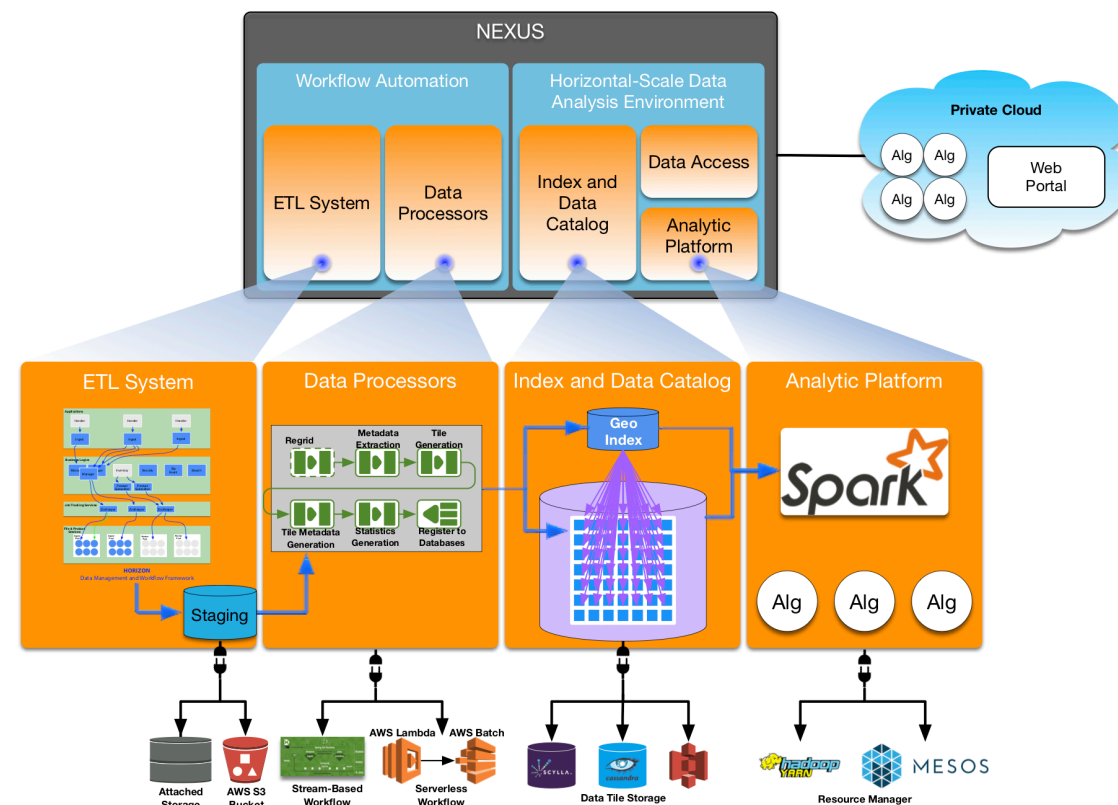
MapReduce Model



Two-Database Architecture

NEXUS' Pluggable Architecture

- **Several container-based deployment options**
 - Local on-premise cluster
 - Private Cloud (OpenStack)
 - Amazon Web Service
- **Automate Data Ingestion with Image Generation**
 - Cluster based
 - Serverless (Amazon Lambda and Batch)
- **Data Store Options**
 - Apache Cassandra
 - ScyllaDB
 - Amazon Simple Storage Service (S3)
- **Resource Management Options**
 - Apache YARN
 - Apache MESOS
- **Analytic Engine Options**
 - Custom Apache Spark Cluster
 - Amazon Elastic MapReduce (EMR)
 - Amazon Athena (work-in-progress)

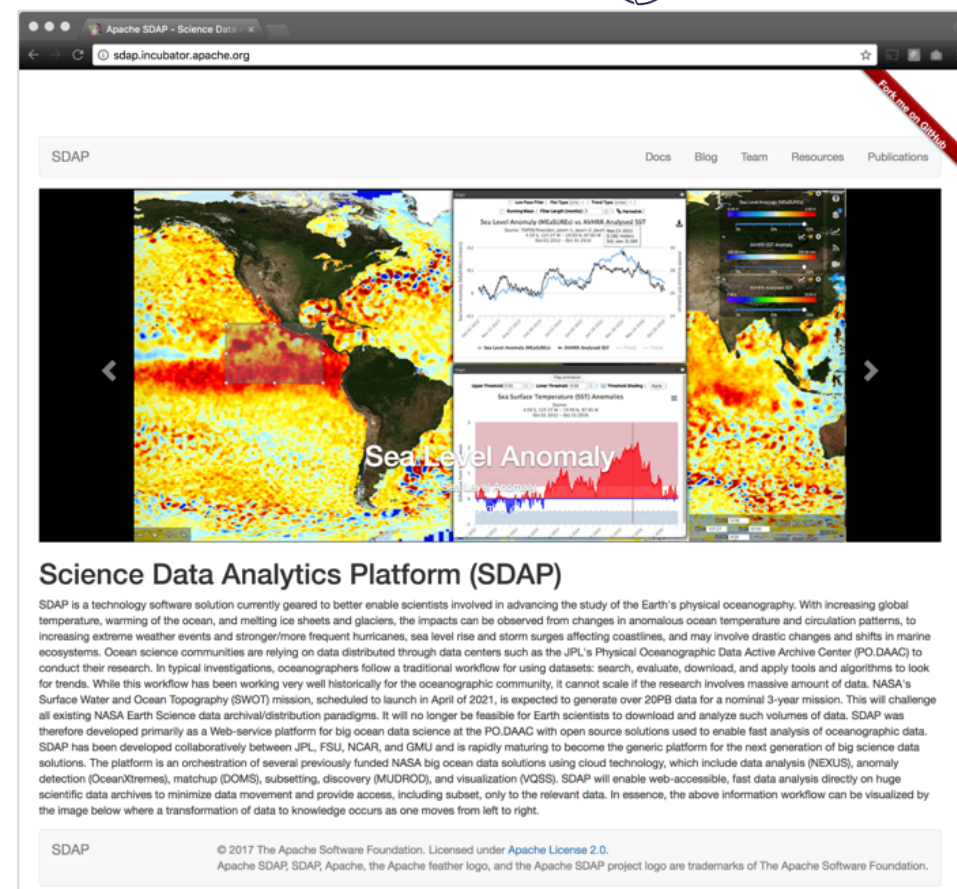


Apache SDAP's NEXUS supports public/private Cloud and local cluster deployments

Free and Open Source Software

- After more than two years of active development, on October 2017 we established Apache Software Foundation and established the **Science Data Analytics Platform (SDAP)** in the **Apache Incubator**
- Technology sharing through Free and Open Source Software (FOSS)
- Why? Further technology evolution that is restricted by projects / missions
- It is more than GitHub
 - Quarterly reporting
 - Reports are open for community review by over 6000 committers
 - SDAP has a group of appointed international mentors
- **SDAP and many of its affiliated projects are now being developed in the open**
 - Support local cluster and cloud computing platform support
 - Fully containerized using Docker and Kubernetes
 - Infrastructure orchestration using Amazon CloudFormation
 - Satellite and model data analysis: time series, correlation map,
 - In situ data analysis and collocation with satellite measurements
 - Fast data subsetting
 - Upload and execute custom parallel analytic algorithms
 - Data services integration architecture
 - OpenSearch and dynamic metadata translation
 - Mining of user interaction and data to enable discovery and recommendations

<http://sdap.apache.org>

Come see us at **ApacheCon North America**
 Las Vegas, Sept 9-12, 2019

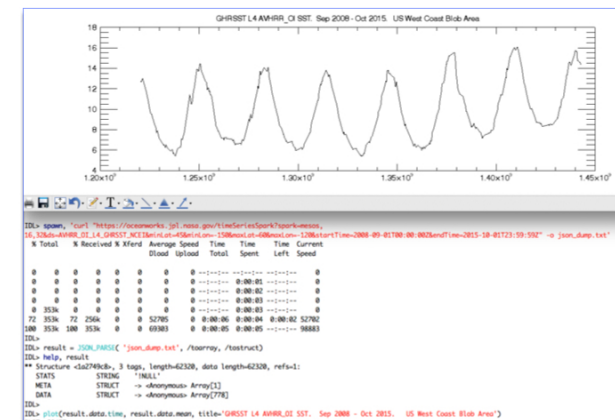
Interact with NEXUS using any Programming Language

```
IDL> spawn, 'curl
"https://oceanworks.jpl.nasa.gov/timeSeriesSpark?spark=mesos,16,32&ds=AVHRR_OI_L4_GH
RSST_NCEI&minLat=45&minLon=-150&maxLat=60&maxLon=-120&startTime=2008-09-
01T00:00:00Z&endTime=2015-10-01T23:59:59Z" -o json_dump.txt'
```

% Total	% Received	% Xferd	Average Speed	Time	Time	Time	Current
			Dload Upload	Total	Spent	Left	Speed
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	353k	0	0	0	0	0	0
72	353k	72	256k	0	52705	0	52702
100	353k	100	353k	0	69303	0	98883

```
IDL>
IDL> result = JSON_PARSE( 'json_dump.txt', /toarray, /tostruct)
IDL> help, result
** Structure <1a2749c8>, 3 tags, length=62320, data length=62320, refs=1:
  STATS      STRING      '!NULL'
  META       STRUCT      -> <Anonymous> Array[1]
  DATA      STRUCT      -> <Anonymous> Array[778]
```

```
IDL>
IDL> plot(result.data.time, result.data.mean, title='GHRSSST L4 AVHRR_OI SST. Sep
2008 - Oct 2015. US West Coast Blob Area')
PLOT <29457>
```



Credit: Ed Armstrong
Jun. 05, 2018


NEXUS integration points

- JPL Oceanworks portal
 - <https://oceanworks.jpl.nasa.gov>
 - Satellite data analysis and in situ matchup
- NASA Sea Level portal
 - <https://sealevel.nasa.gov/data/data-analysis-tool>
 - Altimeter sea level rise and variation
- NASA GRACE portal
 - <https://grace.jpl.nasa.gov/data/data-analysis-tool/>
 - Mass change analysis
- Others in progress.....

NEXUS roadmap into a visualization service

- Integration in the PO.DAAC State Of The Ocean (SOTO) visualization service
 - <https://podaac-tools.jpl.nasa.gov/>
 - SOTO presents long time series global images of ocean temperature, wind, topography, salinity, color, currents and much more
 - But no capability of time series analysis
 - NEXUS integration as the backend of SOTO version 5 provides this capability
 - Currently in a test-user acceptance phase
 - Let's look at some of its functionality.....

SOTO 5.0 Analytics Capability


State of the Ocean 5.0
?
🔗
⚙️
⏪

DATASETS
CHART

Primary Dataset
Sea Surface Temperature (L4, 1km, Daily)
 GHRST MEaSUREs MUR/PO.DAAC

Comparison Dataset
 None Selected


Chart Type
 Time Series

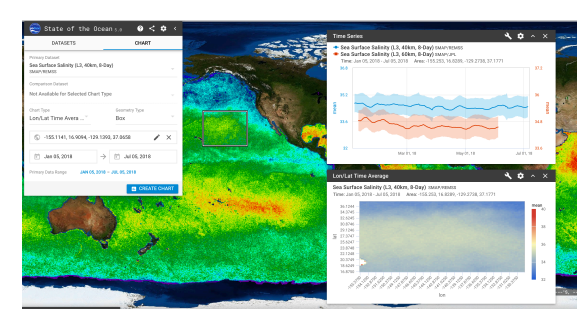
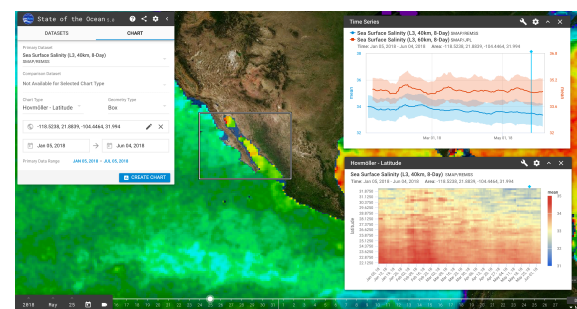
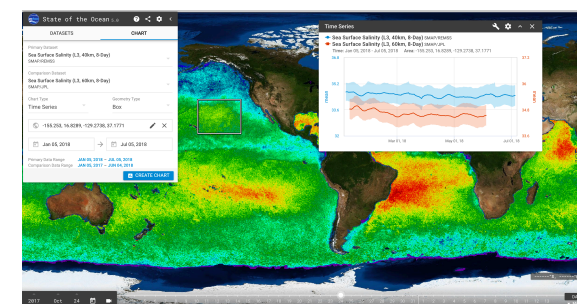
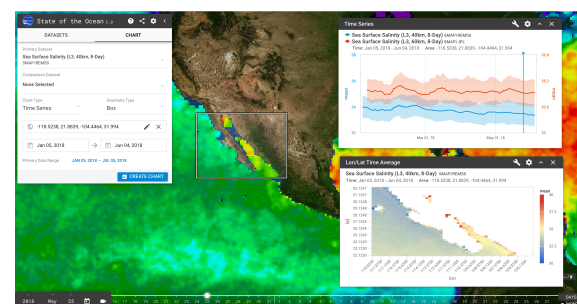
Geometry Type
 Box

📍 -126.1411, 3.9578, -110.2593, 12.4344

📅 Aug 10, 2007 → 📅 Aug 18, 2007

Primary Data Range **AUG 10, 2007 – MAY 20, 2019**

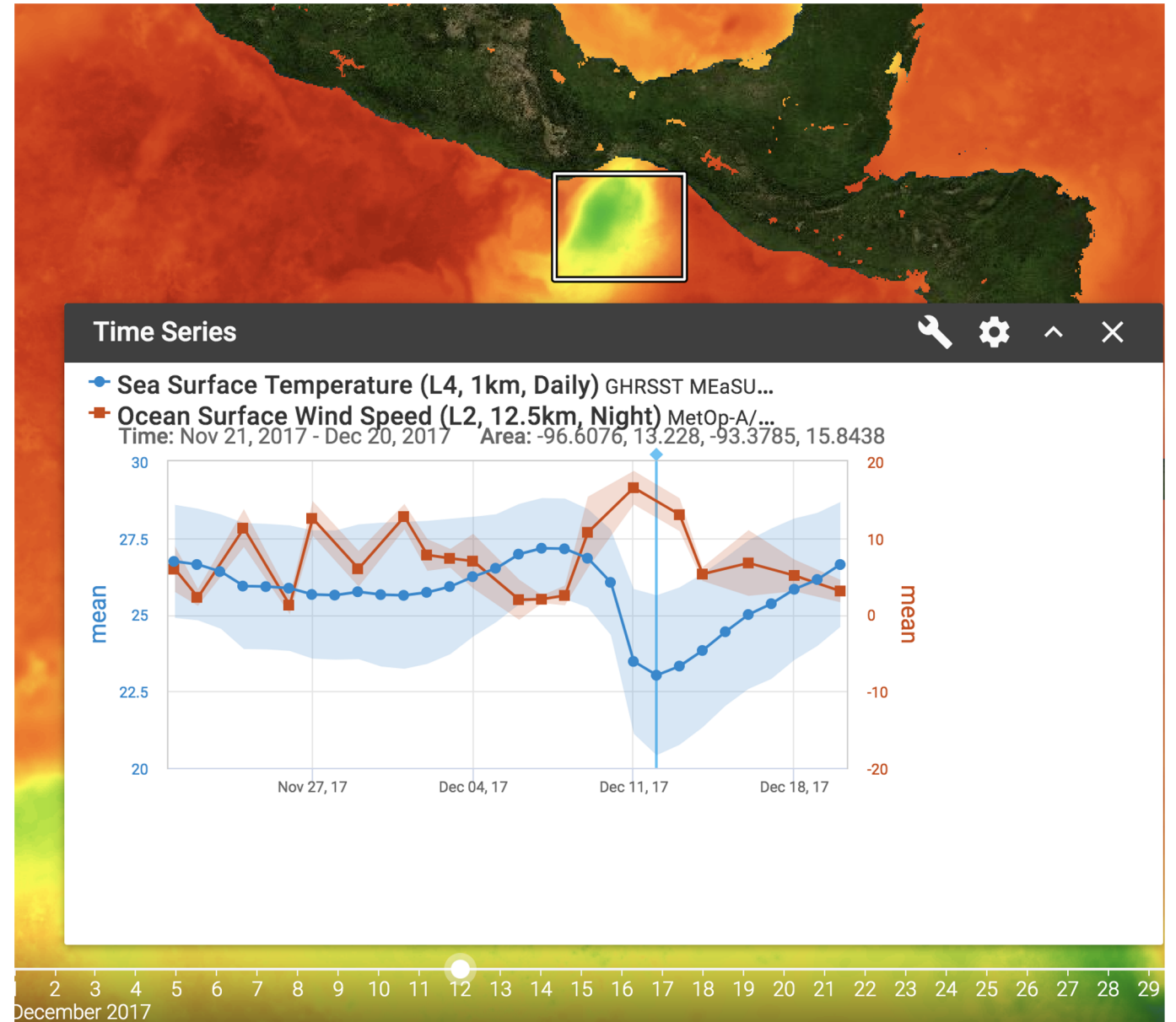
 **CREATE CHART**



Plots: Time Series

- Use Case #1: Tehuantepecers

- Geometry Types:
 - Box
 - Point
- Chart Options:
 - Variables:
 - Min
 - Max
 - Mean
 - Cnt
 - Std
 - Time
 - Display Options
 - Lines and Dots
 - Dots Only
 - Bars



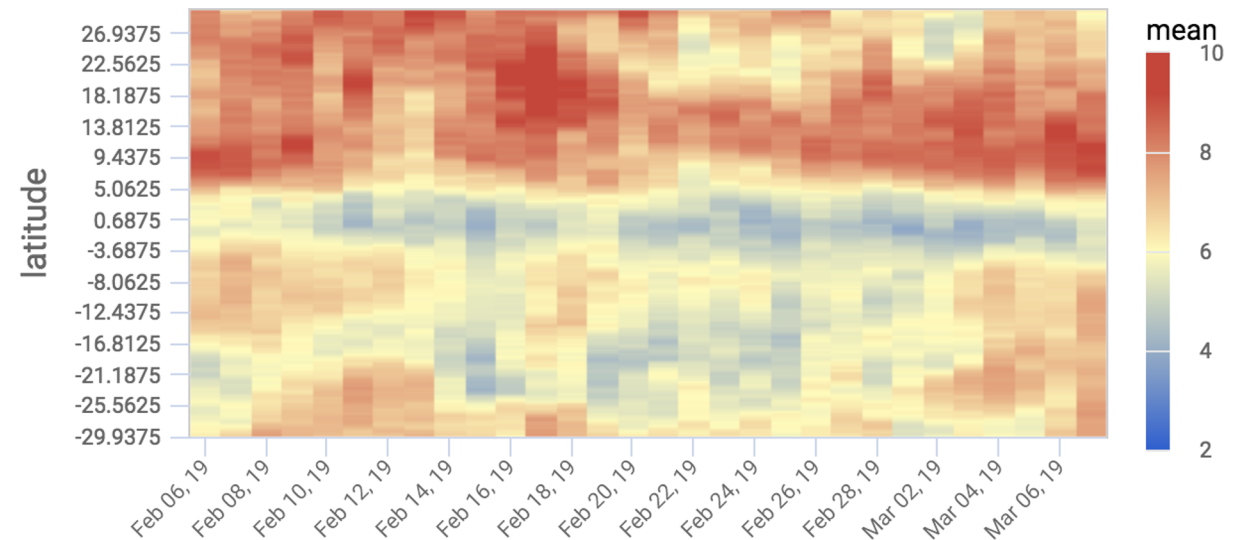
Plots: Hovmoller - Latitude

- Geometry Types:
 - Box
 - Point
- Chart Options:
 - Variables:
 - Min
 - Max
 - Mean
 - Cnt
 - Std
 - Latitude
 - Display Options
 - Invert X/Y

- Use Case #1 - Trade Winds

Ocean Surface Wind Speed (L2, 12.5km, Day) MetOp-A/ASCAT

Time: Feb 06, 2019 - Mar 07, 2019 **Area:** -180, -30, 50, 30



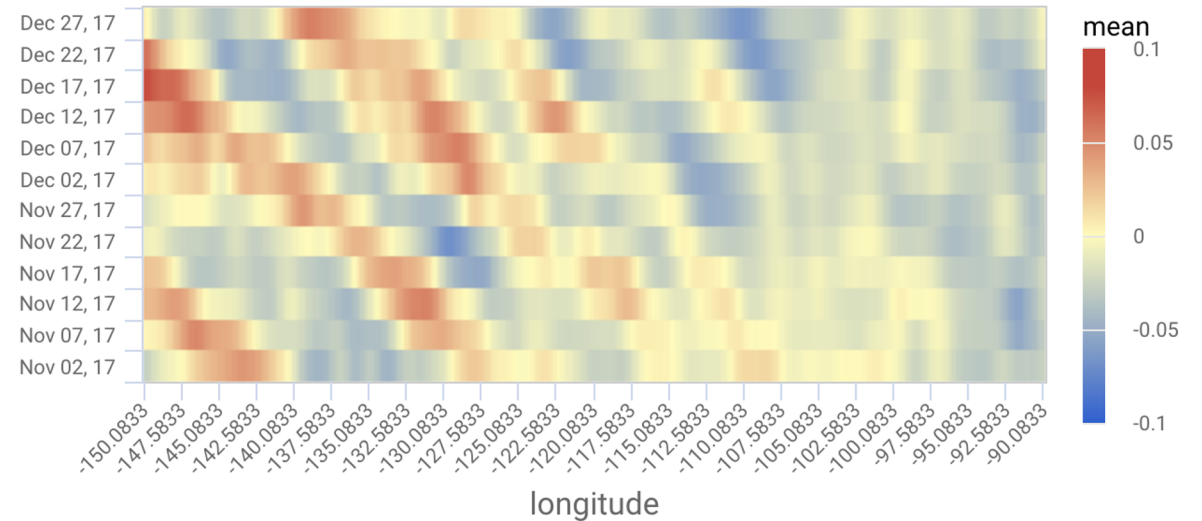
Plots: Hovmoller - Longitude

- Geometry Types:
 - Box
 - Point
- Chart Options:
 - Variables:
 - Min
 - Max
 - Mean
 - Cnt
 - Std
 - Longitude
 - Display Options
 - Invert X/Y

- Use Case #1: [El Nino 3 Box](#)

Sea Surface Height Anomaly (L3, 19km, 5-Day) MEaSURES/JPL

Time: Nov 01, 2017 - Dec 30, 2017 **Area:** -150, -5, -90, 5



Plots: Time-Average (Lat/Lon)

- Geometry Types:

- Box
- Point

- Chart Options:

- Variables:

- Min
- Max
- Mean
- Cnt
- Std
- Longitude

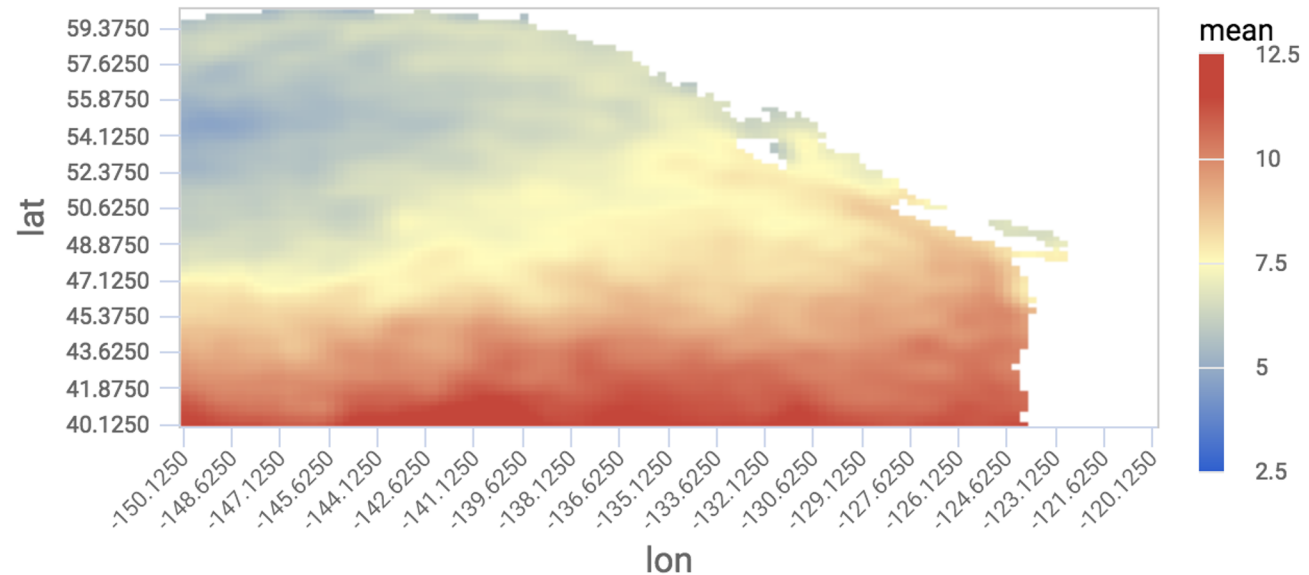
- Display Options

- Invert X/Y

- Use Case #1 - The Blob

Sea Surface Temperature (L4, 25km, Daily) GHRSSST/AVHRR-OI + InSitu

Time: Feb 17, 2019 - Mar 17, 2019 **Area:** -150, 40, -120, 60



Summary

- Increasingly Big Data (gridded and ungridded) will require technologies that *harmonize data, tools, services and computation resources so the scientist can concentrate on results and not on data conditioning or preparation*
- There are a number of technologies including NEXUS to meet this challenge
- NEXUS has demonstrated a successful deployment in the NASA Sea Level and GRACE portals
- Ongoing roadmap to deployment as the SOTO visualization backend
 - Beta testers will be welcomed
- Thanks !

Disclaimer: Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.